DOCUMENT RESUME

ED 455 292 TM 033 146

AUTHOR Abedi, Jamal; Leon, Seth; Mirocha, Jim
TITLE Validity of Standardized Achievement Tests for English Language Learners.
PUB DATE 2001-04-00
NOTE 12p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Correlation; Elementary Secondary Education; *English (Second Language); *Limited English Speaking; *Standardized Tests; *Validity
IDENTIFIERS Iowa Tests of Basic Skills; Latent Variables; Stanford Achievement Tests

ABSTRACT

The concurrent validity of standardized achievement tests (the Stanford 9 and the Iowa Tests of Basic Skills) was examined using data from different school districts nationwide and a latent variable modeling approach. Items in the standardized achievement tests in several content areas were divided into parcels. Parcel scores were used to create latent variables. Students' grade point average, teachers' ratings, and other achievement scores were also used to create external-criterion latent variables. The standardized achievement latent variable was correlated with the external-criterion latent variables. The results suggest that: (1) there is a strong correlation between the standardized achievement and external-criterion latent variables; (2) this relationship is much stronger when latent variables rather than measured variables are used; and (3) the correlation between standardized achievement and external criterion latent variables is significantly larger for the population of students not of limited English proficiency (LEP) than for the LEP population. It is speculated that the low correlation between the two latent variables in the case of the LEP group is due to the impact of language factors. That is, language factors act as construct irrelevant sources. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 455 292

Validity of Standardized Achievement Tests for English Language Learners

Jamal Abedi

Seth Leon

Jim Mirocha

University of California, Los Angles

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Paper presented at the 2001 Annual Meeting of the American Educational Research Association, Seattle, Washington

TM033146

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
J. Abedi
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

BEST COPY AVAILABLE

Perspective

Concurrent validity of standardized achievement tests (Stanford 9 and ITBS) was examined on the data from different school district nationwide using a latent-variable modeling approach. Items in the standardized achievement tests in several content areas were divided into parcels. Parcel scores were used to create latent variables. Students' grade point average, teachers' rating, and other achievement scores were also used to create external-criterion latent variable. Standardized achievement latent variable was correlated with the external-criterion latent variables. The results suggested that: (1) there is a strong correlation between the standardized achievement and external-criterion latent variables; (2) this relationship is much stronger when latent-variables rather than measured variables are used; and (3) the correlation between standardized achievement and external-criterion latent variables is significantly larger for the non-LEP than the LEP population. We speculate that the low correlation between the two latent variables in the case of the LEP group is due to the impact of language factors. That is, language factors act as construct irrelevant sources.

Data Sources

The data for this study were obtained from four locations:

Site 1. Site 1 is a large urban school district. ITBS performance data from 1999 for grades 3 through 8 were obtained. The data included student responses to test items (item-level data), subsection scores, and student background data. These subsection summary scores were grouped into four categories that included math concepts and estimation, math problem solving and data interpretation, math computation, and reading.

Site 2. Site 2 is a state with a very large number of LEP students. Data were obtained on Stanford 9 test for all students in Grades 2 to 11 who were enrolled in the state-wide public schools for the 1997-1998 academic year. These data included student responses to test items (item-level data), subsection scores, and student background data. The background data included gender, ethnicity, free/reduced-price lunch participation, parent education, student LEP status, and Students with Disabilities (SD) status.

Site 3. Site 3 is an urban school district. Stanford 9 test data were available for all students in Grades 10 and 11 for the 1997-1998 academic year. These data

included student responses to test items (item-level data), subsection scores, student background data, and accommodation data.

Site 4. Site 4 is a state with a significant number of English language learners. The Department of Education in this state gave us access to the Stanford 9 summary test data for all students in Grades 3,6,8 and 10 who were enrolled in the state-wide public schools for the 1997-1998 academic year.

Findings

The results of our analyses on the Stanford 9 item-level data that we reported earlier suggested that language factors may introduce another source of measurement error in the measurement model for LEP students. Internal consistency coefficients were lower for LEP students. There were large differences in the performance of LEP and non-LEP students that were apparent especially with respect to the reading items.

Due to the impact of language factors, the intercorrelation between individual test items, the correlation between items and total test score (internal validity coefficient), and the correlation between item score and total test score with the external criteria (students' achievement data) may be different for LEP and non-LEP students. That is, these relationships may be stronger for non-LEP students. To further examine the hypothesis of differences between LEP and non-LEP students on the structural relationship of the test items, a series of confirmatory factor models were created in site 2 and site 3. Fit indices were compared across LEP and non-LEP groups. The results generally indicated that the relationships between individual items, items with the total test score, and items with the external criteria are higher for non-LEP than for LEP students.

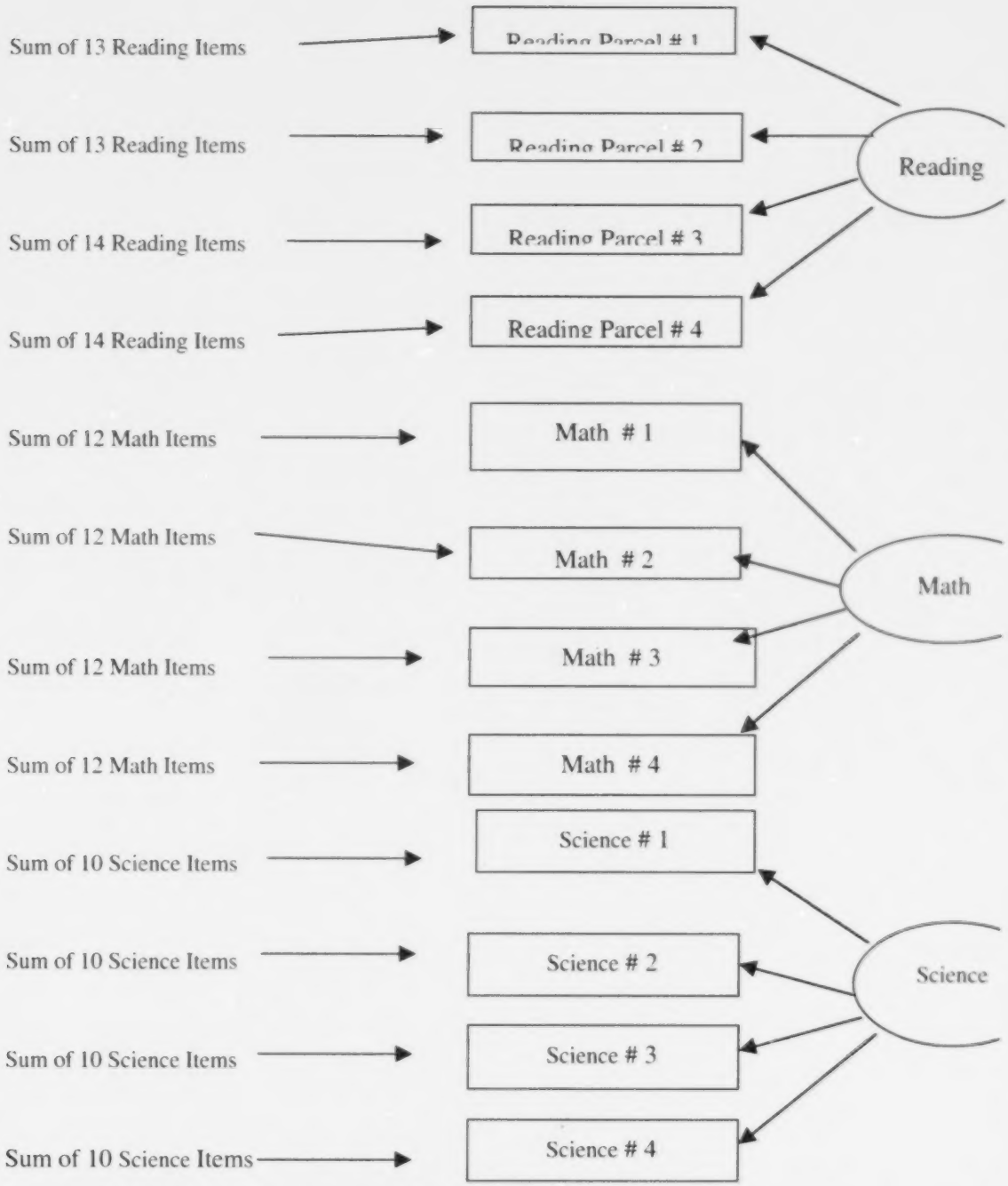


Figure 4.10. Grade 9 Site 1. Simple Structural Equations Model

To compare within-test and cross-test structural relationships between LEP and non-LEP students, a series of simple structure confirmatory models were created. In creating these models, test items in each of the three content areas (reading, science, and math) were grouped as "parcels." Figure 1 presents item-parcels and latent variables for reading, math and science and the correlation between the reading, math and science latent variables for Site 1. As Figure 1 shows, the 52 reading items were grouped into 4 parcels. Each parcel was constructed to systematically contain heterogenous items based on item difficulty. Through this process each parcel contained both easy, difficult and moderately difficult items. The result was a set of homogenous parcels. A reading latent variable was constructed based on these four parcels. Similarly, item parcels and latent variables for science and math were created from the 40 science items and 48 math items though the same process. Correlation between the reading, math and science latent variables were estimated. Models were tested on randomly selected sample populations to demonstrate the consistency of the results.

Table 1 shows the results of the structural models run for grade 9. As data in Table 4.11 show, correlations of item parcels to the latent factors are consistently lower for LEP students than they are for non-LEP students. This finding was true for all parcels regardless of which grade or which sample of the population was tested. For example, in grade 9 for LEP students the correlation for the four reading parcels ranged from a low of .719 to a high of .779 across the two samples as shown in table 4.11. In comparison, for non-LEP students the correlation for the four reading parcels ranged from a low of .832 to a high of .858 across the two samples. The item parcel correlations were also larger for non-LEP students then for LEP students in math and science. Again these results were consistent across the different samples. The paired correlations between the latent factors were also larger for non-LEP students then they were for LEP students. This gap in latent factor correlations between non-LEP and LEP students was especially large when there was a larger language demand difference on the test items. For example, in the grade 9 sample population #1 the correlation between latent factors for math and reading for non-LEP students was .782 compared to just .645 for LEP students. When comparing the latent factor correlations between reading and science from the same population the correlation was still larger for non-LEP students (.837) than for LEP students (.806), but the gap between the correlations decreased. This is likely due to a larger language demand difference between the reading and math tests as compared to the reading and science tests. Multiple group structural models were run to test whether the differences between non-LEP and LEP students

mentioned above were significant. There was significant differences for all constraints tested at the $p<.05$ level.

Table 1. Site 2 Data 1998, Grade 9 Stanford 9 Reading and Math and Science Structural Modeling Results (DF=51)

Factor Loadings	Non-LEP (N=22,782) Sample #1	Non-LEP (N=22,782) Sample #2	LEP (N=4,872) Sample #1	LEP (N=4,872) Sample #2
Reading Comp.				
Parcel 1	.852	.853	.723	.719
Parcel 2	.841	.844	.734	.739
Parcel 3	.835	.832	.766	.779
Parcel 4	.858	.858	.763	.760
Math Factor				
Parcel 1	.818	.821	.704	.699
Parcel 2	.862	.860	.770	.789
Parcel 3	.843	.843	.713	.733
Parcel 4	.797	.796	.657	.674
Science Factor				
Parcel 1	.678	.681	.468	.477
Parcel 2	.679	.676	.534	.531
Parcel 3	.739	.733	.544	.532
Parcel 4	.734	.736	.617	.614
Factor Correlation				
Reading vs Math	.782	.779	.645	.674
Reading vs Science	.837	.839	.806	.802
Science vs Math	.870	.864	.796	.789
Goodness of fit				
Chi Square	488	446	152	158
NFI	.997	.998	.992	.992
NNFI	.997	.997	.993	.993
CFI	.998	.998	.995	.995

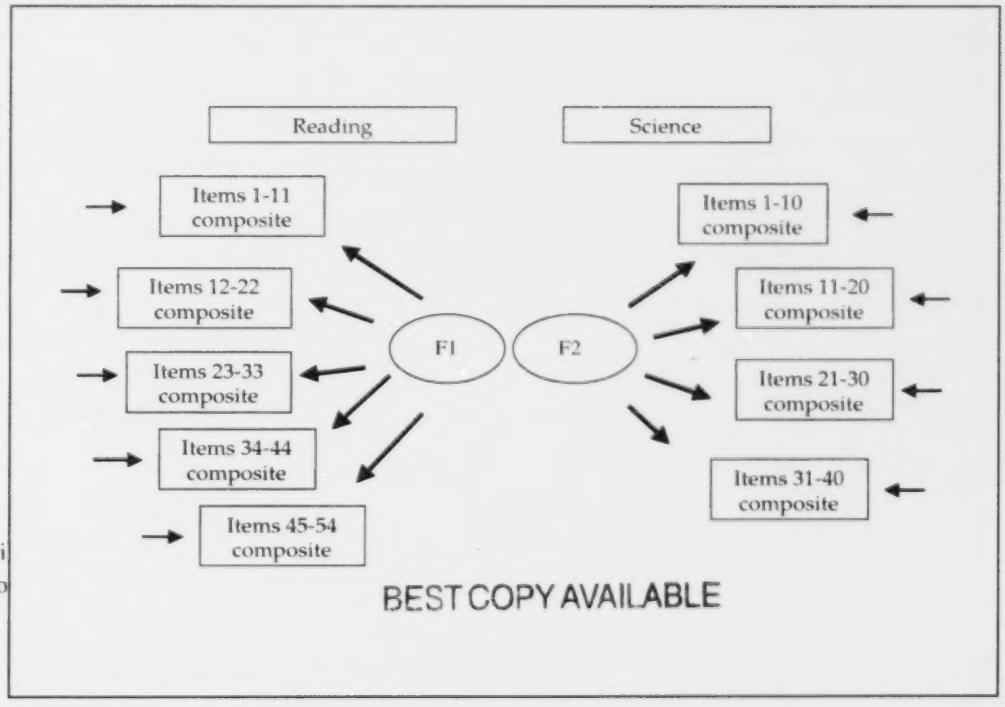
* There was significant invariance for all constraints tested with multiple group model (Non-LEP/LEP).

Site 3 Structural Modeling

To compare within-test and cross-test structural relationships between LEP and non-LEP students, a series of simple structure confirmatory models were created also for site 3. In creating these models, test items in each of the three content areas (reading, science, and math) were grouped as "parcels" Several item-parcels were constructed for each test. Items-parcels were used as measured variables, and one latent variable was created to represent each content area. Correlation coefficients between the content-based latent variables were then estimated.

Reading tests for Grades 10 and 11 had 54 items. Five parcels (measured variables) and a *reading latent variable* based on the five parcels were constructed. Similarly, four parcels and a *science latent variable* were constructed from the 40-item science tests for Grades 10 and 11. A *math latent variable* based on five parcels from the 48-item math tests in Grades 10 and 11 was also created.

Figure 2 presents item-parcels and latent variables for reading and science and the correlation between the reading and science latent variables. As Figure 2 shows, the 54 reading items were grouped into 5 parcels (items 1-11 were grouped into parcel 1, items 12-22 were grouped into parcel 2, and so on). A reading latent variable was constructed based on the five parcels and was labeled as F1. Similarly, 4 parcels were created from the 40 science items and a science latent variable was created (F2). Correlation between the reading and science latent variables was estimated.



Fi
co

BEST COPY AVAILABLE

Table 2 summaries the results of our analyses for the model that was presented in Figure 2 for Grade 10. To do a cross-validation study, we divided the entire population of students into two groups: (a) The group called *even cases* consists of students who were assigned even serial numbers, and (b) the group called *odd cases* consists of students who were assigned odd serial numbers. Because student names were ordered alphabetically, the assignment of subjects to the two groups was considered *systematic random sampling*.

Table 2. Grade 10 Stanford 9 Reading and Science Structural Modeling Results (DF = 24), Site 3 School District

	All cases (N=9,182)	Even cases (N=4,591)	Odd cases (N=4,591)	Non-LEP (N=8,918)	LEP (N=264)
Goodness of fit					
Chi Square	2040	966	1098	1940	106
NFI	.931	.935	.925	.932	.831
NNFI	.897	.904	.890	.899	.792
CFI	.931	.936	.927	.933	.861
Factor Loadings					
Reading Variables					
Composite 1	.687	.695	.679	.685	.628
Composite 2	.692	.698	.687	.687	.697
Composite 3	.745	.738	.751	.741	.724
Composite 4	.822	.823	.821	.823	.712
Composite 5	.689	.688	.691	.691	.550
Science Variables					
Composite 1	.667	.671	.662	.665	.623
Composite 2	.564	.554	.575	.565	.449
Composite 3	.649	.648	.650	.652	.547
Composite 4	.453	.451	.456	.461	.262
Factor Correlation					
Reading vs. Math	.811	.824	.797	.809	.815

Note. NFI = Normed Fit Index. NNFI = Non-Normed Fit Index. CFI = Comparative Fit Index.

In Table 2, we have reported the goodness of fit statistics, correlation coefficients between the items parcels and the latent variables (factor loadings), and the correlation between the two latent variables (reading and math). These statistics

were reported separately for the entire group of students in Grade 10, for the two cross-validation subgroups, and for LEP and non-LEP students. Statistics under the goodness of fit section include Chi-square, Normed Fit Index (NFI), Non-Normed Fit Index (NNFI), and Comparative Fit Index (CFI) (see Bentler, 1992; Bentler, & Bonett, 1980).

As the data in Table 4.14 suggest, the fit statistics for the entire group are very similar to those reported for the cross-validation subgroups (even-cases and odd-cases) and to those reported for the non-LEP groups. For example, the NFI is .931 for the entire group of Grade 10 students. For the even-cases, it is .935; for the odd-cases, it is .925, and for the non-LEP group, it is .932. However, for the LEP group, the NFI drops to .831 which indicates that for LEP students, the fit is not as good as for the non-LEP group or for the entire group. This may be due to the fact that for non-LEP students, the language factor may introduce a new source of bias (measurement error) or construct irrelevant variance as we speculated earlier.

Additionally, Table 4.14 reports correlations between the parcel scores and the reading and science latent variables (factor loadings) for all students in Grade 10, for the two cross-validation groups (even and odd cases), and for the non-LEP and LEP groups. These correlations are very similar for all groups except for the non-LEP group. For the non-LEP group, the correlations are generally lower. For the entire group, for the cross-validation groups and for the non-LEP students, the correlations range from .451 to .823 with an average of .663. For the LEP group, the correlations range from .262 to .724 with an average of .577. These results indicate that the latent models do not provide as strong a structural relationship for the LEP group as for the non-LEP groups. This may be partly due to impact of language factors on the measurement.

Table 4.14 also reports correlation coefficients between the factors (latent variables). These correlations are very similar across the subgroups including the LEP subgroup in this table (Grade 10, reading and math). However, in other cases, these correlations follow the same pattern of lower relationship for LEP students.

Multiple Group Factor Analyses: Testing the Invariance Between Structural Relationship of the LEP and Non-LEP Groups

In the previous sections we reported the results of simple-structure confirmatory factor analyses showing the structural relationship of test scores between LEP/non-LEP across the three content areas. The results of our analyses showed differences on factor loadings and factor correlations between the LEP and non-LEP groups. In additional analyses presented in this section, we created

multiple-group factor models to test the statistical significance of such differences. We examined the hypothesis of invariance of factor loadings and factor correlations between the LEP and non-LEP groups. Specifically, we tested the following null hypotheses:

- Correlations between parcel scores and a reading latent variable are the same for the LEP and non-LEP groups.

- Correlations between parcel scores and a science latent variable are the same for the LEP and non-LEP groups.

- Correlations between parcel scores and a math latent variable are the same for the LEP and non-LEP groups.

- Correlations between content-based latent variables are the same for the LEP and non-LEP groups.

Table 3 summarizes the results of analyses for reading and math tests for students in Grade 10. The data in Table 3 include fit indices for LEP and non-LEP groups, correlations between the parcel scores and the content-based latent variables (factor loadings), and the correlations between the latent variables. Hypotheses regarding the invariance of factor loadings and factor correlations between LEP and non-LEP were tested. Significant differences between the LEP and non-LEP groups at or below .05 nominal levels were identified. These differences are indicated by an asterisk (*) next to each of the constraints. There were several significant differences between the LEP and non-LEP on the correlations between parcel scores and latent variables. For example, on the math subscale, factor loadings between the LEP and non-LEP groups on parcels 2 and 3 were significant. Table 4.18 also shows a significant difference between the LEP and non-LEP on the correlation between reading and math latent variables.

Table 4.18. Grade 10 Stanford 9 Reading and Math Structural Modeling Results (Parcels Ordered by Item Number). Site 3 School District

Goodness of fit	Model #1 (DF=75)		Model #2 (DF=74)	
Chi Square	2938		2019	
NFI	.916		.943	
NNFI	.902		.933	
CFI	.918		.945	
Factor Loadings	Non-LEP (N=8,947)	LEP (N=303)	Non-LEP (N=8,947)	LEP (N=303)
Reading Composite 1	.677	.683	.679	.685
Composite 2	.683	.612	.684	.613
Composite 3	.738	.695	.739	.696
Composite 4	.826	.816	.824	.812
Composite 5	.693	.723	.690	.720
Math Composite 1	.735	.763	.752	.788
Composite 2	.659	.702*	.667	.716*
Composite 3	.623	.730*	.592	.685*
Composite 4	.724	.774	.722	.774
Composite 5	.389	.471	.330	.391
Error Correlation				
E10 vs. E8	---	---	.329	.365*
Factor Correlation				
Reading vs. Math	.719	.624*	.723	.622*

These results indicate that:

- Findings from the two cross-validation samples are very consistent and provide evidence for the validity of analyses.
- Structural models show a better fit for non-LEP than for LEP students.
- Correlations between parcel scores and the content-based latent variables are generally lower for LEP students.
- Correlations between the content-based latent variables are lower for LEP students.
- These results are all indicative of a possible language factor as a source of measurement error for LEP students.[1]

[1] For a complete report of the results of existing data analyses email Jamal Abedi, UCLA/CRESST at: jabedi@cse.ucla.edu or call: (310) 206-4346.

BEST COPY AVAILABLE